THE INVALUABLE GUIDE
TO ALL ASPECTS OF STATISTICS

Oxford

DICTIONARY OF Statistics

GRAHAM UPTON AND IAN COOK

OXFORD
UNIVERSITY PRESS

Great Clarendon Street, Oxford, OX2 6DP,
United Kingdom

Oxford University Press is a department of the University of Oxford.
It furthers the University's objective of excellence in research, scholarship,
and education by publishing worldwide. Oxford is a registered trade mark of
Oxford University Press in the UK and in certain other countries

© Oxford University Press 2002, 2006, 2008, 2014

The moral rights of the authors have been asserted
Database right Oxford University Press (maker)

First published 2002
Second edition 2006
Second edition revised 2008
Third edition 2014

All rights reserved. No part of this publication may be reproduced, stored in
a retrieval system, or transmitted, in any form or by any means, without the
prior permission in writing of Oxford University Press, or as expressly permitted
by law, by licence or under terms agreed with the appropriate reprographics
rights organization. Enquiries concerning reproduction outside the scope of the
above should be sent to the Rights Department, Oxford University Press, at the
address above

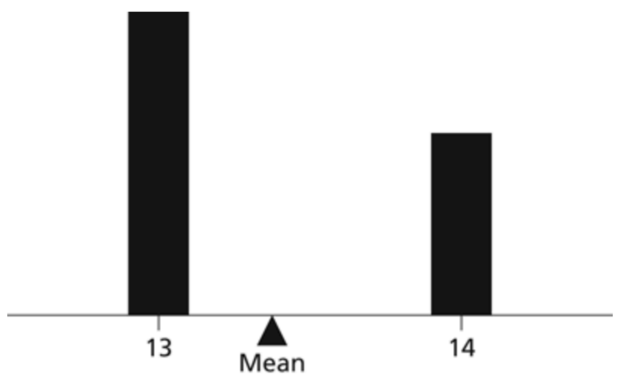
You must not circulate this work in any other form
and you must impose this same condition on any acquirer

British Library Cataloguing in Publication Data
Data available

ISBN 978–0–19–967918–8
ebook ISBN 978–0–19–104463–2

Links to third party websites are provided by Oxford in good faith and
for information only. Oxford disclaims any responsibility for the materials
contained in any third party website referenced in this work.

The mean can be interpreted as the centre of gravity, or centre of mass, of a system of particles of masses $f_1, f_2,\ldots,f_n$ at points $x_1, x_1,\ldots, x_n$.



Sample mean. The data are the numbers of eruptions of the Old Faithful geyser during the first eight days of August 1978. The sample mean is seen to be the balance point of the observations.

sample size The number of *observations in a *sample.

sample space **(universal set)** A complete set of all possible results or **outcomes** for an experiment or observational procedure. The concept was introduced by *von Mises in 1931. The sample space is usually denoted by S or E.

An **event** is a particular collection of outcomes, and is a subset of the sample space. For example, when a die is thrown and the score observed, the sample space is {1, 2, 3, 4, 5, 6}, and a possible event is 'the score is even' i.e. {2, 4, 6}. If all the possible outcomes are equally likely, then the probability of an event A is given by

$$\mathrm{P}(A) = \frac{\text{Number of events in subset of sample space corresponding to } A}{\text{Number of events in sample space}}.$$

The word 'event' was used in this context by *de Moivre in 1718.

The subset of the sample space for the event 'the score is both even and odd' is an example of the **empty set**, usually denoted by ϕ, and $\mathrm{P}(\phi)=0$.

The subset of the sample space for the event 'the score is less than 10', is the whole sample space, S, and $\mathrm{P}(S)=1$.

See also BOOLEAN ALGEBRA; COMPLEMENTARY EVENT; INTERSECTION; UNION.

sample standard deviation The square root of the *sample variance.

sample variance A measure of the variability of a set of *data. For data $x_1, x_2,\ldots, x_n$, with *sample mean

$$\frac{(|a-b|-1)^2}{a+b},$$

which, if the drugs are really equally effective, may be taken to be an observation from a *chi-squared distribution with one *degree of freedom. The generalization to more than two matched samples is provided by the **Cochran Q test**.

MD-plot *See* BLAND-ALTMAN PLOT.

mean Familiarly known as the *average, the word 'mean' is used as a shorthand for either the *population mean or the *sample mean, depending on context. *See also* EXPECTED VALUE.

mean absolute deviation **(MAD)** A *measure of spread. For *observations x_1, $x_2, \ldots,$ x_n, with *sample mean $\bar{x}$ and *median m, the mean absolute deviation about the mean is

$$\frac{1}{n}\sum_{j=1}^{n}|x_j-\bar{x}|,$$

and the mean absolute deviation about the median is

$$\frac{1}{n}\sum_{j=1}^{n}|x_j-m|.$$

If X and Y are completely unrelated (i.e. are *independent) then $\rho=0$. If $\rho=0$ then X and Y are said to be *uncorrelated variables. However, ρ is concerned only with linear relationships, and the fact that $\rho=0$ does not imply that X and Y are independent.

population covariance For two *random variables X and Y, this is the difference between the *expected value of their product and the product of their separate expected values. It is denoted by Cov (X,Y):

$$\mathrm{Cov}(X,Y)=\mathrm{E}(XY)-\mathrm{E}(X)\times\mathrm{E}(Y).$$

If X and Y are *independent then $\mathrm{Cov}(X, Y)=0$. However, if $\mathrm{Cov}(X, Y)=0$ then X and Y may not be independent. A useful result is $\mathrm{Var}(aX+bY)=a^2\mathrm{Var}(X)+2ab\ \mathrm{Cov}(X, Y)+b^2\mathrm{Var}(Y)$, where Var denotes *variance, and a and b are constants. The term 'covariance' was used by Sir Ronald *Fisher in 1930. *See also* POPULATION CORRELATION COEFFICIENT.

population mean The *average value of some *variable that is measured for all members of a (possibly infinite) population. If the value of the variable for a randomly chosen member of the

population is denoted by X, then the population mean is the *expected value of X and is usually denoted by μ (the notation μ was introduced in 1936 by Sir Ronald *Fisher in the sixth edition of his *Statistical Methods for Research Workers*). Similarly, the **population variance**, usually denoted by σ^2, is the mean of the squared differences between the values of the members of the population and the population mean: this is the expected value of $(X-\mu)^2$.

population median In a *population of values of the *variable X, the population median, m, is a value for which $\mathrm{P}(X\geq m)=\mathrm{P}(X\leq m)$.

population parameter A key quantity that determines the precise shape of a *distribution. For example, the shape of a *Poisson distribution is determined by the *parameter λ, and that of a *normal distribution by the parameters μ and σ.

population pyramid A diagram (*see overleaf*) for representing the age distribution of a population. It is really a *histogram in which age is plotted vertically and *frequency, or relative frequency (i.e. *proportion), is plotted horizontally. Often drawn as a back-to-back pyramid with one side for males and the

hypothesis if $|r|$ is too large. *See also* COEFFICIENT OF DETERMINATION; RANK CORRELATION COEFFICIENT.

http://www.stat.tamu.edu/~west/ph/coreye.html

- Applet.

sample covariance Given the n pairs of *observations $(x_1, y_1), \ldots, (x_n, y_n)$, the sample covariance, c, is given by

$$c = \frac{1}{n-1}\left\{\sum_{j=1}^{n} x_j y_j - \frac{1}{n}\left(\sum_{j=1}^{n} x_j\right)\left(\sum_{j=1}^{n} y_j\right)\right\}.$$

sample distribution function The equivalent of the *distribution function for a *sample of *data. Let the ordered data be $x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}$; then the sample distribution function $F_n(x)$ is given by

$$F_n(x) = \begin{cases} 0 & x < x_{(1)}, \\ j/n & x_{(j)} \le x < x_{(j+1)}, \quad 1 \le j \le (n-1), \\ 1 & x_{(n)} \le x. \end{cases}$$

sample mean The sample mean (or, simply, 'the mean') of a set of n items of *data $x_1, x_2, \ldots, x_n$ is $\left(\sum_{j=1}^{n} x_j\right)$, which is the arithmetic average of the numbers $x_1, x_2, \ldots, x_n$. The mean is usually denoted by placing a bar over the symbol for the variable being measured. If the variable is x the mean is denoted by $\bar{x}$. If the data constitute a *sample from a *population, then the sample mean is an unbiased estimate of the *population mean.

For example, the numbers of eruptions of the *Old Faithful geyser during the first eight days of August 1978 were 13, 13, 13, 14, 14, 14, 13, and 13. The mean is

$$(13+13+13+14+14+14+13+13)/8 = 13.375.$$

If the data are collected in *frequency form so that values $x_1, x_2, \ldots, x_n$ are obtained with frequencies $f_1, f_2, \ldots, f_n$ the mean is

$$\frac{\sum_{j=1}^{n} f_j x_j}{\sum_{j=1}^{n} f_j}.$$

For example, with the eruption data there are just two values, $x_1 = 13$ and $x_2 = 14$. Their respective frequencies are $f_1 = 5$ and $f_2 = 3$, so the mean is $\{(5 \times 13) + (3 \times 14)\}/(3+5) = 13.375$.

If the data are grouped into classes with mid-values $x_1, x_2, \ldots, x_c$ and corresponding *class frequencies $f_1, f_2, \ldots, f_c$, an approximate value for the mean of the original data is the **grouped mean**

$$\frac{\sum_{j=1}^{c} f_j x_j}{\sum_{j=1}^{c} f_j}.$$

A Dictionary of
Statistics

THIRD EDITION

GRAHAM UPTON
and IAN COOK

OXFORD
UNIVERSITY PRESS